Journal of Biomolecular Techniques • Volume 33(3); 2022 Sep

Commercial ChIP-Seq Library Preparation Kits Performed Differently for Different Classes of Protein Targets

MS Simper[1] L Della Coletta[1] S Gaddis[1] K Lin[1] CD Mikulec[1] Y Takata[1] MW Tomida[1] D Zhang[1,2] DG Tang[1,3] MR Estecio[1] J Shen[4,1,5] Yue Lu[1]

[1]Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Science Park, Smithville, Texas 78957, USA,

[2]Present Address: College of Biology, Hunan University, Changsha 410082, China,

[3]Present Address: Department of Pharmacology and Therapeutics, Roswell Park Cancer Institute, Buffalo, New York 14263, USA,

[4]Department of Epigenetics and Molecular Carcinogenesis,

[5]and Program in Genetics and Epigenetics, MD Anderson Cancer Center UT Health Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center, Smithville, Texas 78957, USA

Association of Biomolecular Resource Facilities

Published on: Nov 14, 2022

URL: https://jbt.pubpub.org/pub/b0nisrdj

ABSTRACT

Background: Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) is a powerful method commonly used to study global protein–DNA interactions including both transcription factors and histone modifications. We have found that the choice of ChIP-Seq library preparation protocol plays an important role in overall ChIP-Seq data quality. However, very few studies have compared ChIP-Seq libraries prepared by different protocols using multiple targets and a broad range of input DNA levels. **Results:** In this study, we evaluated the performance of 4 ChIP-Seq library preparation protocols (New England Biolabs [NEB] NEBNext Ultra II, Roche KAPA HyperPrep, Diagenode MicroPlex, and Bioo [now PerkinElmer] NEXTflex) on 3 target proteins, chosen to represent the 3 typical signal enrichment patterns in ChIP-Seq experiments: sharp peaks (H3K4me3), broad domains (H3K27me3), and punctate peaks with a protein binding motif (CTCF). We also tested a broad range of different input DNA levels from 0.10 to 10 ng for H3K4me3 and H3K27me3 experiments. **Conclusions:** Our results suggest that the NEB protocol may be better for preparing H3K4me3 (and potentially other histone modifications with sharp peak enrichment) libraries; the Bioo protocol may be better for preparing H3K27me3 (and potentially other histone modifications with broad domain enrichment) libraries, and the Diagenode protocol may be better for preparing CTCF (and potentially other transcription factors with well-defined binding motifs) libraries. For ChIP-Seq experiments using novel targets without a known signal enrichment pattern, the NEB protocol might be the best choice, as it performed well for each of the 3 targets we tested across a wide array of input DNA levels.

ADDRESS CORRESPONDENCE TO: Jianjun Shen, Department of Epigenetics and Molecular Carcinogenesis; Program in Genetics and Epigenetics, MD Anderson Cancer Center UT Health Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center, Smithville, TX 78957, USA (E-mail: jianjunshen888@gmail.com).

ADDRESS CORRESPONDENCE TO: Yue Lu, Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Science Park, Smithville, TX 78957, USA (Phone: (713) 563-7960; E-mail: YLu4@mdanderson.org).

MS Simper and L Della Coletta contributed equally to this work.

Conflict of Interest Disclosures: The authors declare that they have no competing interests.

Keywords: next generation sequencing, ChIP-Seq, library preparation, quality control

INTRODUCTION

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) has become the method of choice for studying global protein–DNA interactions and chemical modifications of histone proteins. Since

Keji Zhao and colleagues first coined the term “ChIP-Seq” in 2007,[1] more than 6,500 publications have employed ChIP-Seq methods to 1) survey interactions between protein and DNA to provide insight into mechanisms central to biological processes and disease states; 2) identify the genomic locations of chromatin-associated proteins; and 3) identify posttranslational modifications affecting histones, chromatin-modifying complexes, and other chromatin-associated proteins. Many factors must be considered and procedures optimized to complete successful ChIP-Seq experiments. These include, but are not limited to, the amount of tissue and/or number of cells available; the degree to which proteins are crosslinked to chromatin by formaldehyde; the method of chromatin fragmentation (eg, sonication or enzymatic cleavage); the precision of fragment size selection; the quality of the antibody selected for immunoprecipitation; the chosen library preparation protocol; the cluster generation and sequencing platform; and post-sequencing data quality checks and subsequent data processing. To establish proper standards for the ChIP-Seq research community, the ENCODE and modENCODE consortia set guidelines for many of these factors, especially for validating antibodies, replicating experiments, required sequencing depth, and scoring and evaluating ChIP-Seq data.[2]

Although many factors can introduce technical bias affecting next-generation sequencing (NGS) outcomes, the main sources of bias stem from chromatin structure, PCR amplification, and read-mapping effects.[3] With over 10 years of operation as a next-generation core facility serving 120 laboratories, our experience indicates that the ChIP-Seq library preparation protocol plays an important role in overall ChIP-Seq outcomes, particularly when using ultra-low levels of input DNA. Smaller amounts of template require more PCR cycles to generate enough material for sequencing, but amplification bias increases with each PCR cycle.[3] In cases with limited numbers of cells and/or amounts of input DNA, increased numbers of unmapped reads and PCR duplications are common.[4] Over the past several years, many commercial kits have become available for ChIP-Seq library preparation, including kits developed for use with limited numbers of cells and/or ultralow DNA input. Sundaram et al. compared 7 commercial and/or homemade ChIP-Seq library preparation methods at 2 input DNA levels (1 ng and 0.1 ng) using a PCR-free dataset as a reference.[5] Each kit was evaluated by performing ChIP-Seq for H3K4me3 and then comparing unmapped reads, PCR amplification–derived duplicates, reproducibility, and sensitivity and specificity relative to the PCR-free reference dataset.[5] They found that the Swift Biosciences Accel-NGS 2S ChIP-Seq library preparation method performed the best, the Rubicon Genomics ThruPLEX kit performed the second best, and the Sigma-Aldrich SeqPlex method performed poorly.

Earlier studies provided the research community with valuable information; however, they evaluated only a single target[4],[5],[6] and included neither commonly used ChIP-Seq library preparation kits (eg, Next Ultra II kit from NEB and KAPA HyperPrep kit from Roche) nor kits specificially designed for low-input samples (eg, the Diagenode kit for low input). The Bioo NEXTflex ChIP-Seq kit (PerkinElmer) was included in our evaluation because a number of our customers used its earlier version and generated good quality data. Here, we compared the performance of these 4 ChIP-Seq library preparation protocols (NEB, KAPA, Diagenode, and Bioo) on 3 separate targets: histone H3 trimethylated on lysine 4 (H3K4me3), histone H3 trimethylated on

lysine 27 (H3K27me3), and the transcription factor (TF) CCCTC binding factor (CTCF). We carefully chose these targets to represent the three typical outcomes seen in ChIP-Seq experiments: targets that display sharp peaks, as expected for H3K4me3; targets that display broad peaks, as expected for H3K27me3; and targets that display discrete, punctate peaks, as expected for proteins that bind to specific DNA sequences such as CTCF.[6] While CTCF is not representative of all TF given its abundant number of sites, it is an example of narrow peak binding. To test a broad range of input levels, we used 6 different amounts of input DNA, ranging from 0.1 to 10 ng, for H3K4me3 and H3K27me3 ChIP-Seq. All of our ChIP-Seq libraries were evaluated with respect to sequencing library complexity, reproducibility, and specific quality metrics suitable for the enrichment patterns of the 3 different protein targets. Our study indicates that the ChIP-Seq library preparation protocols performed differently for different classes of protein targets. The NEB protocol may be the best choice for H3K4me3 (and potentially other histone modifications with sharp peak enrichment); the Bioo protocol may be the best choice for H3K27me3 (and potentially other histone modifications with broad domain enrichment), though not at very low DNA levels; and the Diagenode protocol may be the best choice for CTCF (and potentially other TFs that bind to specific DNA sequence motifs). For ChIP-Seq experiments that target proteins with unknown signal enrichment patterns, the NEB protocol might be a good choice, as it performed well with all 3 targets, and the NEB libraries behaved consistently across different input DNA levels.

MATERIALS AND METHODS

Cell culture and fixation

The androgen-sensitive human prostate adenocarcinoma cell line LNCaP was purchased from the American Type Culture Collection (ATCC) and cultured in Roswell Park Memorial Institute (RPMI) 1640 (+) L-Glutamine medium from Gibco Life Technologies (Thermo Fisher Scientific, Waltham, Massachusetts) supplemented with 10% fetal bovine serum, 100 units/mL penicillin-streptomycin, and 1 mM sodium pyruvate. This line was authenticated regularly in our institutional CCSG Cell Line Characterization Core and also examined to be free of mycoplasma contamination. Cells were maintained in 5% CO_2 at 37°C and cultured in 10-cm plates to a confluency of 70-80% before fixation. LNCaP cells were fixed in 1% methanol-free formaldehyde (Thermo Fisher Scientific) in RPMI for 10 min at room temperature followed by quenching for 5 min in 125 mM glycine (Sigma, St. Louis, Missouri) with low-speed shaking on an orbital platform. Plates were washed 2× with ice-cold phosphate-buffered saline (PBS) (pH 7.4) (Thermo Fisher Scientific) to eliminate any residual media, and 7 mL cold PBS containing 1× Complete Protease Inhibitors Cocktail, ethylenediamine tetraacetic acid (EDTA) free (Roche, Basel, Switzerland) were immediately added. Plates were kept on ice while the cells were harvested by scraping. Cells were transferred to 15-mL conical tubes (2 plates/tube) and collected by centrifugation (4 min at 805 × *g* at 4°C). Cells were immediately used for chromatin preparation.

Chromatin preparation

Fixed LNCaP cell pellets were resuspended by pipetting in sodium dodecyl sulfate (SDS) lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris pH 8.1, and 1X Complete Protease Inhibitors Cocktail [Roche]) and incubated for at least 10 min on ice. Three hundred milliliters of SDS lysis buffer were added for every 2-3 million cells, and 300 mL of cell lysates were transferred to a 1.5-mL tube from Diagenode, Inc. (Denville, New Jersey) for sonication. Sonication was performed using a Diagenode Bioruptor Plus with 22 cycles of 30 s high power (on) followed by 30 s rest (off) at 4°C. After the intial 22 sonication cycles, the samples were allowed to rest on ice for 15 min and then subjected to an additional 22 cycles of sonication using the same conditions. This shearing protocol routinely resulted in a chromatin preparation with fragments sized between 200 bp and 700 bp. The cell lysates were cleared by centrifugation (10 min at 14,500 × *g* at 4°C), and each supernatant was transferred to a new tube. Two microliters of the supernatant were diluted with 180 mL SDS lysis buffer and used for reverse crosslinking. The DNA was purified using the QIAquick PCR Purification kit protocol (Qiagen, Hilden, Germany) and quantified using a Qubit fluorometer, and 1 ng of purified DNA was loaded into an Agilent 2100 Bioanalyzer (Santa Clara, California) using High Sensitivity DNA reagents to ensure that the desired size profile was obtained.

Chromatin immunoprecipitation

The following antibodies were used for chromatin immunoprecipitation: anti-histone H3 [ab1791] from Abcam (Cambridge, United Kingdom) and anti-H3K4me3 [17-614], anti-H3K27me3 [17-622], and anti-CTCF [07-729] from MilliporeSigma (St. Louis, Missouri). For each immunoprecipitation, 15 mL of Dynabeads Protein A magnetic beads (Invitrogen, Carlsbad, California) were combined with 15 mL Dynabeads Protein G magnetic beads (Invitrogen) and washed 3× in blocking solution (1× PBS/0.5% bovine serum albumin). After the final wash, the Dynabeads were resuspended in 250 mL blocking solution, and 10 mg antibody was added for each 15-30 mg chromatin. The mixture was rotated overnight at 4°C and washed 3× with blocking solution before resuspension in 100 mL blocking solution. The chromatin lysate was diluted 1:10 using ChIP dilution buffer (16.7 mM Tris-HCl [pH 8.1], 167 mM NaCl, 1.2 mM EDTA, 0.01% SDS, and 1.1% Triton X-100) with protease inhibitors, and 100 mL of diluted lysate were removed and saved as Input for comparison during analysis. One milliliter diluted lysate was then added to the antibody/bead mixture and incubated overnight on a rotator at 4°C. A magnet was used to collect the beads, which were then washed 3× with 1 mL radioimmunoprecitation assay (RIPA) washing buffer (Teknova, Hollister, California). Each supernatant was removed and transferred to a fresh tube. A final wash was performed with 1 mL 1× TE (Promega, Madison, Wisconsin) containing 50 mM NaCl, and the beads were collected by centrifugation (1 min at 960 × *g* at 4°C) before resuspension in 110 mL SDS lysis buffer. Crosslinking was reversed, and DNA was purified from both the immunoprecipitated and the input samples before analysis using a Bioanalyzer. To validate the ChIP experiments prior to library preparation and sequencing, vendor-provided primers were used for qPCR as follows: hGAPDH-promoter region-specific primers for H3K4me3, hAlpha satellite control primers for H3K27me3, and hSCN4A primers for CTCF.

NEBNext Ultra II DNA Library Prep Kit for Illumina

ChIP-Seq libraries were prepared using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, Inc., Ipswich, Massachusetts) following the manufacturer's protocol. Briefly, ng (1, 5, and 10 ng) and sub-ng (100, 250, and 500 pg) amounts of fragmented DNA were used to repair the ends before ligation to a NEBNext Adapter for Illumina sequencing. The adapter-ligated DNA was then enriched using PCR (1 cycle at 98°C for 30 s followed by a specific number of cycles depending on the amount of input DNA [Table 1] at 98°C for 10 s and 65°C for 75 s and then 1 cycle at 65°C for 5 min]. DNA was purified using AMPure XP beads (Beckman Coulter, Brea, California) after adapter ligation and PCR enrichment.

Table 1

PCR amplification cycle number specific to kit and input levels

Input DNA (ng)	Number of PCR amplification cycles			
	KAPA	Diagenode	Bioo	NEB
10	7	7	11	6
5	9	8	13	7
1	13	11	15	9
0.5	15	13	17	10
0.25	17	14	19	11
0.1	20	16	22	13

KAPA HyperPrep Kit

ChIP libraries were prepared using the KAPA HyperPrep Kit (Roche) following the manufacturer's protocol. Briefly, ng (1, 5, and 10 ng) and sub-ng (100, 250, and 500 pg) amounts of fragmented DNA were used to repair the ends before ligation to a diluted NEXTflex DNA adapter (PerkinElmer, Waltham, Massachusetts). The adapter-ligated DNA was then enriched using PCR (1 cycle at 98°C for 45 s followed by a specific number of cycles depending on amount of input DNA [Table 1] at 98°C for 15 s, 60°C for 30 s, and 72°C for 30 s and then 1 cycle at 72°C for 1 min). DNA was purified using AMPure XP beads (Beckman Coulter) after adapter ligation and PCR enrichment.

Diagenode MicroPlex Library Preparation Kit v2

ChIP libraries were prepared using the Diagenode MicroPlex Library Preparation Kit v2 (Diagenode) following the manufacturer's protocol. Briefly, ng (1, 5, and 10 ng) and sub-ng (100, 250, and 500 pg) amounts of DNA were repaired to create blunt ends for ligation to a MicroPlex stem-loop adapters. The adapter-ligated DNA was then enriched and indexed using PCR (1 cycle at 72°C for 3 min; 1 cycle at 85°C for 2 min; 1 cycle at 98°C for 2 min; and 4 cycles at 98°C for 20 s, 67°C for 20 s, and 72°C for 40 s followed by a specific number of cycles depending on the amount of DNA input [see Table 1] at 98°C for 20 s and 72°C for 50 s). The total number of PCR cycles listed in Table 1 does not include the 4 cycles of stage 1 and is just the number of stage 2 cycles. DNA was purified using AMPure XP beads (Beckman Coulter) after adapter ligation and PCR enrichment.

NEXTflex ChIP-Seq Library Prep Kit for Illumina Sequencing (Bioo)

ChIP libraries were prepared using the NEXTflex ChIP-Seq Kit (Bioo Scientific, now PerkinElmer) following the manufacturer's protocol. Briefly, ng (1, 5, and 10 ng) and sub-ng (100, 250, and 500 pg) amounts of fragmented DNA were end repaired and adenylated prior to ligation to a NEXTflex ChIP adapter. The adapter-ligated DNA was then enriched using PCR (1 cycle at 98°C for 2 min followed by a specific number of cycles depending on amount of DNA input [Table 1] at 98°C for 30 s, 65°C for 30 s, and 72°C for 1 min and then 1 cycle at 72°C for 4 min). DNA was purified using AMPure XP beads (Beckman Coulter) after repair, adapter ligation, and PCR enrichment.

RNA-Seq library preparation

Libraries were prepared using the Illumina TruSeq Stranded Total RNA (Cat. #RS-122-2301) kit according to the manufacturer's protocol, starting with 1 microgram total RNA as previously described.[7] Briefly, ribosomal (rRNA) rRNA-depleted RNAs were fragmented and converted to complementary DNA (cDNA) with reverse transcriptase. The resulting cDNAs were converted to double-stranded cDNAs and subjected to end repair, A-tailing, and adapter ligation. The constructed libraries were amplified using 8 cycles of PCR.

Sequencing

Each ChIP-Seq library was checked for quality using a 2200 TapeStation (Agilent Technologies). A KAPA Library Quantification Kit (Roche) was used to quantify the libraries for pooling, and a final concentration of 1.5 nM was loaded onto an Illumina cBot for cluster generation before sequencing with an Illumina HiSeq 3000 using a single-read 50 bp run.

Bioinformatics analysis

Mapping

In order to make meaningful comparisons with our own existing data (data not shown), all ChIP-Seq reads were trimmed to 36 bp and mapped to the human genome (hg19). Mapping was performed using Bowtie (version 1.1.2),[8] allowing for no more than 2 mismatches and retaining only reads that were mapped to unique positions. About 91-96% of reads were mapped to the human genome, with 78-82% being uniquely mapped. To avoid PCR bias, when multiple reads were mapped to the same genomic position, only 1 read was retained for analysis.

Peak/domain calling

H3K4me3 and CTCF peaks were detected using MACS (version 1.4.2, *P* value cutoff 1e-5 and window size 300 bp). Peaks that overlapped ENCODE blacklisted regions were removed. To avoid any possible effects of total H3 library quality, H3K4me3 peak calling was performed without a control library reference. CTCF peaks in each library were called by taking the corresponding total input library as a control. The H3K27me3-enriched domains were initially identified using the enriched domain detector (EDD version 1.1.16) without any significance threshold (ie, the false discovery rate (FDR) cutoff for EDD was set to be larger than 1). The EDD gap penalty and bin size were set to 10 and 20 kb, respectively. After peak calling with EDD, for each 20 kb bin, the *z*-score was calculated as:

$(\hat{p} - p)/\sigma$, where $\hat{p}$ = *Number of ChIP reads/(Number of ChIP reads + Number of total H3 reads)*, $p = 0.5$ and

$$\sigma = \sqrt{p \times (1-p)/(Number\ of\ ChIP\ reads + Number\ of\ total\ H3\ reads)}$$

The *P* value for each candidate domain was calculated using Stouffer's *z*-score method by combining all the bins within the putative domain. The *P* values were corrected using the method of Benjamini and Hochberg. The domains with FDR ≤ 0.05 were called as significantly enriched with H3K27me3. To minimize the effect of total H3 library quality, a combined total H3 was used as a control for the detection of H3K27me3-enriched domains with 15 M reads from the total H3 library prepared by each protocol at the 10-ng level.

Signal landscape

For the H3K4me3 and CTCF libraries, each read was extended by 150 bp (ie, the expected average fragment size) to its 3' end. The number of reads mapping to each genomic position was normalized to a total of 10 M mapped reads, averaged over every 10-bp window, and displayed in the UCSC Genome Browser (http://genome.ucsc.edu/).[9] For each H3K27me3 library, the log2 ratio of the number of reads in each 20-kb window over the combined total H3 following normalization, based on the total number of mapped reads, was calculated and displayed in the UCSC Genome Browser.

Gene annotation

Genes from GENCODE Release 29[10] were used to annotate H3K4me3 peaks and H3K27me3-enriched domains. The promoter region was defined as -1,000 bp to +500 bp of a transcription start site (TSS).

RNA-Seq analysis

For gene expression data, two RNA-Seq libraries were generated from LNCaP cells and sequenced (2 x 75 bp paired-end protocol) using an Illumina HiSeq 2500 instrument. Although it was the same cell line, the RNA-Seq was performed at a different time from a different set of cells. Each pair of reads represents a cDNA fragment from the library. The reads were mapped to the human genome (hg19) using TopHat (version 2.0.10).[11] The number of fragments corresponding to each known gene in GENCODE (Release 29) was enumerated using htseq-count (HTSeq package version 0.6.0).[12] Genes shorter than 200 bp and those coding for rRNAs were removed prior to analysis. The FPKM (number of fragments per kilobase per million fragments) value for each gene was calculated and averaged over the 2 replicates.

Quality of peak calling for H3K4me3

The quality of peak calling for H3K4me3 was measured as the percentage of peaks in promoter regions versus the percentage of expressed genes marked by peaks. The promoter region was defined as -1,000 bp to +500 bp of a TSS. To identify the expressed genes, we plotted log10(FPKM) values for all of the genes in a histogram, which revealed a bimodal shape. The histogram was fitted to a density curve using the density function in R. Genes with FPKM values larger than the point of minimal density between the 2 density peaks were defined as expressed genes.

H3K4me3 aggregated signal around TSS/enhancer

The 3,000 bases 5’ and 3’ of a TSS as well as the 3,000 bases 5’ and 3’ from the center of an enhancer were subdivided into 100-bp bins. For each H3K4me3 library, the RPKM (reads per kilobase per million mapped reads) value was calculated for each bin and averaged over all TSSs or enhancers. Enhancers were defined as DNaseI hypersensitive sites that are separated by at least 10 kb from a TSS. DNaseI hypersensitivity data was obtained from the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeUwDnase).

Correlation of H3K27me3 signal with gene expression

For each library, the log2 ratio was calculated from the number of H3K27me3 reads within -2,000 bp to +1,000 bp of each gene’s TSS over the number of total H3 reads. Log2 ratio values versus gene expression values (based on FPKM) were compared using Spearman’s correlation. The TSS of the longest transcript was used to represent each gene’s TSS.

Correlation of H3K27me3 signal with chromatin accessibility

For each library, the log2 ratio was calculated from the number of H3K27me3 reads in each 20-kb window over the number of total H3 reads. Log2 ratio values versus chromatin accessibility scores were compared in 20-kb windows using Spearman's correlation. Chromatin accessibility scores were calculated as RPKM values based on DNaseI-Seq data for LNCaP cells downloaded from the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeUwDnase). As H3K27me3 is typically enriched in facultative heterochromatin,[13] windows lacking a TSS for a gene were excluded from the analysis.

Identification of CTCF motifs in CTCF peaks

The identification of canonical CTCF motifs near CTCF peak summits used the FIMO[14] component of the MEME suite (version 4.10.2)[15] with default parameter settings. The canonical CTCF motif was defined as in JASPAR (Matrix ID MA0139.1).[16]

Reproducibility curves

For comparing H3K4me3 peaks for each pair of H3K4me3 libraries, the peaks were merged and ranked from the strongest to the weakest based on RPKM values. The percentage of peaks common to both libraries from the same number of top peaks was calculated. A similar approach was applied for comparing H3K27me3-enriched 20-kb windows except that the windows were ranked by *z*-scores.

Number of reads used

For all comparisons, the same number of reads was used for all libraries shown on a single plot; however, the number of reads per plot varies across the plots. The number of reads used in each plot is summarized in Table S2.

RESULTS

Experimental design and data quality metrics

The experimental design used for testing the four ChIP-Seq library preparation protocols (Bioo, Diagenode, KAPA, and NEB) is outlined in Figure 1. ChIP DNAs from LNCaP cells were immunoprecipitated with H3K4me3, H3K27me3, and CTCF antibodies before library preparation. These antibodies have been widely used for ChIP.[17],[18],[19],[20],[21],[22] For the H3K4me3 and H3K27me3 ChIP-Seq analyses, we tested 6 different amounts of input DNA for each protocol: 10, 5, and 1 ng and 500, 250, and 100 pg. CTCF DNA was undetectable by fluorometry Qubit following immunoprecipitation; therefore, multiple CTCF immunoprecipitations were combined and then divided equally across the 4 protocols to normalize the amount of input DNA used for each. To ensure reproducibility, 2 independent sets of ChIP-Seq experiments were performed to assess the H3K4me3 (H3K4me3 sets 1 and 2) and H3K27me3 (H3K27me3 sets 1 and 2)

libraries, and 3 independent sets of ChIP-Seq experiments were performed to assess the CTCF libraries (CTCF sets 1-3). All libraries were sequenced on an Illumina HiSeq 3000 using a single-read 50-bp run (see Table S1 for the number of reads generated and mapping rate for each library and Figure 2 for screenshots of signal landscape and peaks called for each library). To assess our ChIP-Seq libraries, we evaluated library complexity for all libraries and reproducibility for the H3K4me3 and H3K27me3 libraries. Based on the known enrichment patterns for our 3 protein targets, we employed different quality metrics for each: H3K4me3 libraries were evaluated by the fraction of reads in peaks (FRiP), the fraction of reads in promoter regions, and the precision and sensitivity of peak calling; H3K27me3 libraries were evaluated by their correlation with gene expression and chromatin accessibility; and CTCF libraries were evaluated by the number of peaks, the percentage of peaks that contained the CTCF motif, and the distance between the CTCF motif and the peak summit. To avoid batch effect, we made the decision of evaluating reproducibility by comparing neighboring input amounts rather than between the 2 sets (sets 1 and 2 for both K3K4me3 and H3K27me3) because the 2 sets were performed more than 1 year apart. Since the results for the sets 2 of H3K4me3 and H3K27me3 generally agree with the results for sets 1, the data for sets 2 are not shown but can be available on request.

Library complexity

Library complexity is commonly used to evaluate the quality of ChIP-Seq libraries.[2],[3],[23] Low complexity libraries often result from insufficient amounts of starting DNA or overamplification by PCR and generate less useful data than do high complexity libraries. To evaluate library complexity, we measured the NRF (nonredundant fraction), defined as the ratio of the number of positions in the genome to which uniquely mappable reads and the total number of uniquely mapped reads (Figure 3).[2] The Bioo libraries yielded the lowest overall complexity, especially at lower amounts of input DNA for both H3K4me3 (Figure 3*A*) and H3K27me3 (Figure 3*B*). As expected, library complexity decreased with lower amounts of input DNA. The Bioo libraries were the most variable in terms of NRFs, whereas the NEB libraries were the most consistent across all input amounts. Notably, at the 100-pg level, the KAPA libraries showed a much more dramatic drop in complexity than did the Diagenode and NEB libraries, with the exception of H3K4me3 set 2. The results were similar when assessing the CTCF libraries (Figure 3*C*). Again, Bioo typically generated the lowest complexity library, NEB generated the highest complexity library, and KAPA performance was variable at lower input amounts.

H3K4me3 ChIP-Seq

Signal portion

To evaluate the portion of true H3K4me3 signal in our libraries, we plotted the FRiP versus sequencing depth at different DNA inputs (Figure 4). Given that H3K4me3 is enriched near TSSs,[1] we also plotted the fraction of reads in promoter regions at each DNA input (Figure S1). Of the tested protocols, the NEB protocol worked well at all DNA input levels and outperformed the other protocols at the 10 and 5 ng levels, and the Diagenode

protocol resulted in the best signal at the picogram DNA levels. The Bioo libraries generated the lowest signal at almost all input levels except the 100 pg level in H3K4me3 set 1, for which the KAPA libraries generated the lowest signal (Figure 4*A* and S1*A*). Surprisingly, although libraries prepared using higher levels of input DNA were expected to have better quality, a general trend within each protocol indicated that the portion of H3K4me3 signal in the library increased when the amount of input DNA decreased (Figures Figure 4*B* and S1*B*). The finding that H3K4me3 signal increased as DNA input decreased was also seen both in the aggregated signal curves around TSSs (Figure S2) and by visual inspection of the signal landscape (Figure 2*A*). This inverse relationship is consistent with a prior study showing H3K4me3 signal over TSSs decreased with increasing amounts of input DNA.[5] Regardless, the H3K4me3 signal from the NEB libraries was the most consistent across all input levels.

Peak calling

H3K4me3 is a common histone modification that generally correlates with the promoters of transcriptionally active genes.[1] To evaluate the quality of peak calling for H3K4me3 from our libraries, we used the percentage of H3K4me3 peaks in promoter regions as a measurement of precision, and we used the percentage of expressed genes marked by those peaks as a measurement of sensitivity (Figure 5). Overall, Bioo libraries produced the lowest percentages for both precision and sensitivity across all DNA inputs (Figure 5*A*). The differences between the Bioo libraries compared to the others became more obvious with each decrease in input DNA. Peak calling was similar across the other 3 protocols at all DNA inputs except that the KAPA libraries were inferior at the 100-pg level in H3K4me3 set 1. The quality of peak calling, at different amounts of input DNA within single protocols (Figure 5*B*), was clearly decreased at the 250-pg and 100-pg levels in H3K4me3 sets 1 and 2 for the Bioo libraries and at 100 pg in H3K4me3 set 1 for the KAPA libraries but remained stable across all inputs for the Diagenode and NEB libraries.

H3K27me3 ChIP-Seq

Correlation with gene expression

H3K27me3 is a histone mark typically associated with gene repression.[1],[13],[24] We evaluated H3K27me3 ChIP by comparing this mark across the genome in our ChIP-Seq data set to our own RNA-Seq gene expression data for this cell line (LNCaP) (Figure 6). As expected, H3K27me3 correlated negatively with gene expression in each of our libraries. However, while the KAPA libraries in H3K27me3 set 2 gave rise to the strongest negative correlation at 5 of the 6 input DNA levels (250 pg-10 ng), KAPA set 1 libraries had the weakest negative correlation with gene expression regardless of DNA input. This exhibits the inconsistency in the KAPA kit performance between libraries prepared at different times. Among the remaining 3 protocols, Bioo libraries had the best negative correlation between the presence of H3K27me3 and gene expression except for the 100-pg library in H3K27me3 set 1. NEB was better than or similar to Diagenode at all input DNA levels (Figure 6*A*). As expected, H3K27me3 ChIP-Seq libraries prepared from higher amounts of input

DNA generally had better negative correlation with gene expression than did ChIP-Seq using lower amounts of DNA (Figure 6*B*).

Correlation with chromatin accessibility

As the H3K27me3 histone mark is associated with heterochromatin,[13],[24] we also evaluated the correlation of H3K27me3 signal in our ChIP-Seq libraries with the signal of chromatin accessibility in LNCaP cells as determined by DNaseI-Seq data available through the UCSC genome browser (Figure 7). Overall, the negative correlation between H3K27me3 signal versus chromatin accessibility agreed with the negative correlation with gene expression. Among the libraries we tested, KAPA libraries were the least inconsistent between H3K27me3 set 1 and set 2 when comparing H3K27me3 signal to chromatin accessibility. Among the remaining 3 protocols, Bioo libraries had the best negative correlation between H3K27me3 signal and chromatin accessibility except for the 100-pg library in H3K27me3 set 1. NEB libraries always performed better than Diagenode libraries at all input DNA levels (Figure 7*A*). As expected, H3K27me3 ChIP-Seq using higher levels of input DNA generally had better negative correlation with chromatin accessibility than did ChIP-Seq using lower levels of input DNA (Figure 7*B*).

CTCF ChIP-Seq

To evaluate the performance of the 4 protocols for determining the genome-wide distribution of the transcription factor CTCF, we assessed the number of peaks called, the percentage of peaks that contained the canonical CTCF binding motif, and the distance between the CTCF motif and the peak summit (Figure 8). Overall, the Diagenode libraries always yielded a greater number of called peaks, and the NEB libraries yielded a lesser number of peaks for all CTCF sets (Figure 8*A*). With respect to identifying a CTCF motif within a peak, the Bioo libraries had the highest percentage of peaks lacking a CTCF motif, whereas the Diagenode libraries had the highest percentage of peaks containing a CTCF motif (Figure 8*B*). Consistently, the offset between a CTCF motif and the peak summit was widest for the Bioo libraries and narrowest for the Diagenode libraries (Figure 8*C*). These findings indicate that the Diagenode libraries might more accurately capture the genome-wide distribution of CTCF, and potentially other TFs, whereas the Bioo libraries might less accurately reflect the true distribution of CTCF.

Reproducibility

To evaluate the reproducibility of the ChIP-Seq protocols, we calculated either the percentage of top peaks (H3K4me3) or enriched 20-kb windows (H3K27me3) shared between two libraries created from the same protocol but differing by an incremental change in the amount of input DNA (ie, 10 ng versus 5 ng, 5 ng versus 1 ng, 1 ng versus 500 pg, 500 pg versus 250 pg, and 250 pg versus 100 pg) and then compared these across protocols (Figure 9). As observed in many of our analyses, the KAPA libraries were inconsistent in set 1 and set 2 for both H3K4me3 and especially H3K27me3. Of the 3 remaining protocols, overall, Bioo libraries had the lowest reproducibility, and the Diagenode libraries had the highest reproducibility. When we compared

reproducibility between all incremental changes in DNA input for each individual protocol, we found that H3K4me3 and H3K27me3 libraries prepared at lower DNA inputs tended to have higher reproducibity (Figure 10). This may be because low input libraries captured only strong signals, which usually have high reproducibility.

DISCUSSION

Study design

In this study, we evaluated the performance of 4 ChIP-Seq library preparation protocols (Bioo, Diagenode, KAPA, and NEB) on 3 separate targets to represent the 3 typical signal enrichment patterns in ChIP-Seq experiments: sharp peaks (H3K4me3), broad domains (H3K27me3), and punctate peaks over sequences bearing a protein-binding motif (CTCF). We also tested a broad range of different input DNA levels ranging from 0.10 to 10 ng for creating our H3K4me3 and H3K27me3 ChIP-Seq libraries. To ensure the reproducibility of our results, 2 independent sets of ChIP-Seq experiments were carried out for H3K4me3 and H3K27me3, and 3 independent sets of ChIP-Seq experiments were carried out for CTCF. To the best of our knowledge, our study is the most comprehensive evaluation of ChIP-Seq protocols across a wide range of input DNA amounts for 3 distinct protein targets associated with 3 different types of signal enrichment pattern in ChIP-Seq. Previous studies were either based on a single target or a smaller range of input DNA levels.[4],[5], [6]

To assess our ChIP-Seq libraries, we evaluated library complexity for all libraries and reproducibility for the H3K4me3 and H3K27me3 libraries. Based on the known enrichment patterns for our 3 protein targets, we employed different quality metrics for each: H3K4me3 libraries were evaluated by the FRiP, the fraction of reads in promoter regions, and the precision and sensitivity of peak calling; H3K27me3 libraries were evaluated by their correlation with gene expression and chromatin accessibility; and CTCF libraries were evaluated by the number of peaks, the percentage of peaks that contained the CTCF motif, and the distance between the CTCF motif and the peak summit.

Protocol performance

An important lesson from our study is that no single protocol consistently outperformed the others for all protein targets across all quality measurements. Overall, the Bioo libraries had lower library complexity, lower reproducibility, and lower performance than the others for preparation of the H3K4me3 and CTCF libraries. However, Bioo libraries performed better than the others for H3K27me3 libraries. Comparatively, the NEB libraries generally had higher library complexity and better H3K4me3 signal, especially at ng DNA levels, and were generally the most consistent across different DNA levels for several metrics (ie, library complexity, signal portion in H3K4me3 libraries, and reproducibility). The Diagenode protocol performed well for preparing CTCF libraries, with a good number of peaks called, the highest percentage of peaks with the CTCF motif, and the shortest distance between the peak summit and CTCF motif. The KAPA libraries were more

variable and inconsistent compared to the others, particularly for library complexity at low input DNA levels, the signal portion in H3K4me3 libraries prepared at low input DNA levels, the correlation between H3K27me3 signal and gene expression/chromatin accessibility, the number of CTCF peaks, and reproducibility. A summary of each protocol's performance based on specific quality metrics is shown in Table 2.

Table 2

Performance of Library Preparation Protocols[*]

	Overall Library Complexity		H3K4me3 ChIP		H3K27me3 ChIP		CTCF ChIP			Reproducibility
	DNA input Level (ng)	DNA Input Level (pg)	Signal Portion	Peak Calling	vs. gene Expression	vs. Chromatin Accessibility	Number of Peaks	% of Peaks with CTCP motif	Distance between Peak Summit and CTCP motif	
Bioo	+	+	+	+	+++	+++	inconsistent	+	+	+
Diagenode	++	++	+++ at pg levels	+++	+	+	+++	+++	+++	+++
KAPA	++	inconsistent	++	++	inconsistant	inconsistent	inconsistant	++	++	inconsistant
NEB	+++	+++	+++ at ng levels	+++	++	++	++	++	++	++

*Ranked from lowest (+) to highest (+++) performance for the indicated parameters

In summary, our study indicates that commercial library preparation kits performed differently for different classes of protein targets. For example, the NEB protocol may be the best choice for H3K4me3 (and potentially other histone modifications with sharp peak enrichment), Bioo may be the best choice for H3K27me3 (and potentially other histone modifications with broad domain enrichment) though not at very low DNA levels, and Diagenode may be the best choice for CTCF (and potentially other TFs that bind a specific DNA-binding motif). For ChIP-Seq experiments that target proteins with unknown signal enrichment patterns, the NEB protocol might be a good choice, as it performed well with all 3 targets, and the NEB libraries produced consistent results across different input DNA levels.

Performance of different input DNA levels and potential problems of some quality metrics

Although libraries prepared using higher levels of input DNA were expected to have better quality (because of more representative DNA material and fewer PCR cycles), only the following metrics increased with increasing DNA input: library complexity, peak calling precision and sensitivity for H3K4me3, and correlation versus gene expression/chromatin sensitivity for H3K27me3. In contrast, the FRiP and the fraction of reads in promoter regions for H3K4me3 and reproducibility for both H3K4me3 and H3K27me3 decreased with increasing DNA input. A previous study from Sundaram et al. also reported an inverse correlation between H3K4me3 signal over TSSs and DNA input.[6] To further investigate why some metrics inversely correlated with input DNA level, we generated scatter plots of H3K4me3 signal intensity in peaks for each input level versus the 10-ng level (Figure S3). Our results suggested that the signal from libraries created with lower amounts of DNA was skewed toward stronger peaks and may explain why the calculated signal portion in low input libraries tended to be higher than in high input libraries. Among all protocols, NEB was less skewed toward strong peaks. Given that the NEB protocol also gave the highest portion of H3K4me3 signal in libraries at 10 and 5 ng, we also compared the other protocols to the NEB protocol at 10 and 5 ng levels using scatter plots of H3K4me3 signal intensity in peaks (Figure S4). Interestingly, we found that instead of being skewed toward strong peaks, the signal from the NEB libraries increased more evenly compared to the others. That is, the signal in weak peaks was also stronger. Together, our data indicate that some metrics such as FRiP that are commonly used to evaluate the quality of ChIP-Seq may not always be reliable. It is possible that samples with better on-target specificity, such as the 10-ng level samples in our study, receive a worse value because the sample captures real but weak signals that are incorrectly classified as background noise. Indeed, in contrast to the H3K4me3 signal around TSSs, higher input DNA levels tended to have stronger signals than lower input levels around enhancers (Figure S5), where H3K4me3 is weakly enriched.[1],[25] Based on our investigation, we suggest examining multiple metrics to determine ChIP-Seq library quality and include an additional visual verification such as a signal intensity scatter plot as used in this study.

In summary, our study confirmed the common belief that the libraries prepared using higher levels of input DNA have a better quality.

Effect of PCR cycles on our study

In this study, different numbers of PCR cycles were used to amplify DNA fragments in ChIP-Seq libraries according to the manufacturer's recommendation, with Bioo requiring the most cycles followed by Diagenode, KAPA, and NEB (Table 1). We chose to follow the manufacturer's recommendations, as we anticipate that is what most users would do and therefore makes our results more meaningful for their experiments. However, PCR introduces bias that can affect the metrics we used to evaluate the ChIP-Seq library quality. For example, library complexity was directly impacted by the number of PCR cycles and correlated perfectly with the number of PCR cycles. On the other hand, none of the other quality metrics we used had a perfect correlation

with the number of PCR cycles. Although PCR is regarded as the most important cause of guanine-cytosine (GC) bias[26] and GC content generally increases with cycle number,[27] the Bioo libraries, which underwent the highest number of PCR cycles, tended to have lower GC content (Figure S6). Our results suggest that although the quality of the ChIP-Seq library can be influenced by the number of PCR cycles, it is more influenced by other factors intrinsic to each specific protocol.

Limitations and future work

There are limitations to this study that could be addressed in the future. For example, here we only investigated the representative protein targets of 2 major regulatory mechanisms: TFs and posttranslational modifications of histones. In the future, we may investigate the third major mechanism: higher order chromatin organization, which might be addressed by targeting nuclear lamina proteins.[28] Lamina-associated domains range from 80 kb to 30 Mb in human fibroblasts. These domains are even broader than those associated with H3K27me3; thus, the protocols' performance on lamina-associated domains may be different from our current study. Future work should also include using paired-end sequencing data to help determine whether the preference to a range of fragment size is a factor influencing the protocol performance.

ACKNOWLEDGMENTS

The authors thank Dr. Briana Dennehey for editorial assistance, Dr. Sharon Dent and Dr. Michael MacLeod for their critical reading of the manuscript, Laura Denton for the preparation of the manuscript, and Joi Holcomb for her help with the preparation of Figure 1.

Author Contributions: JS, YL, MSS, and LDC conceived the project, designed experiments, interpreted data and wrote the manuscript; MSS, LDC, CDM, YT, MWT, and DZ performed experiments; YL, SG, and KL conducted bioinformatics analyses; ME and DGT helped with data interpretation; and SG helped with writing the manuscript. All authors read and approved the manuscript.

Funding/Support: This project was supported by Cancer Prevention and Research Institute of Texas Core Facility Support Awards (RP120348 and RP170002) to JS. The funding agency had no role in the design, collection, analysis, interpretation, and the writing of the manuscript.

Availability of Data and Materials: The raw datasets for the ChIP-Seq protocols have been deposited in Gene Expression Omnibus (GEO) and can be accessed by the accession number GSE196402.

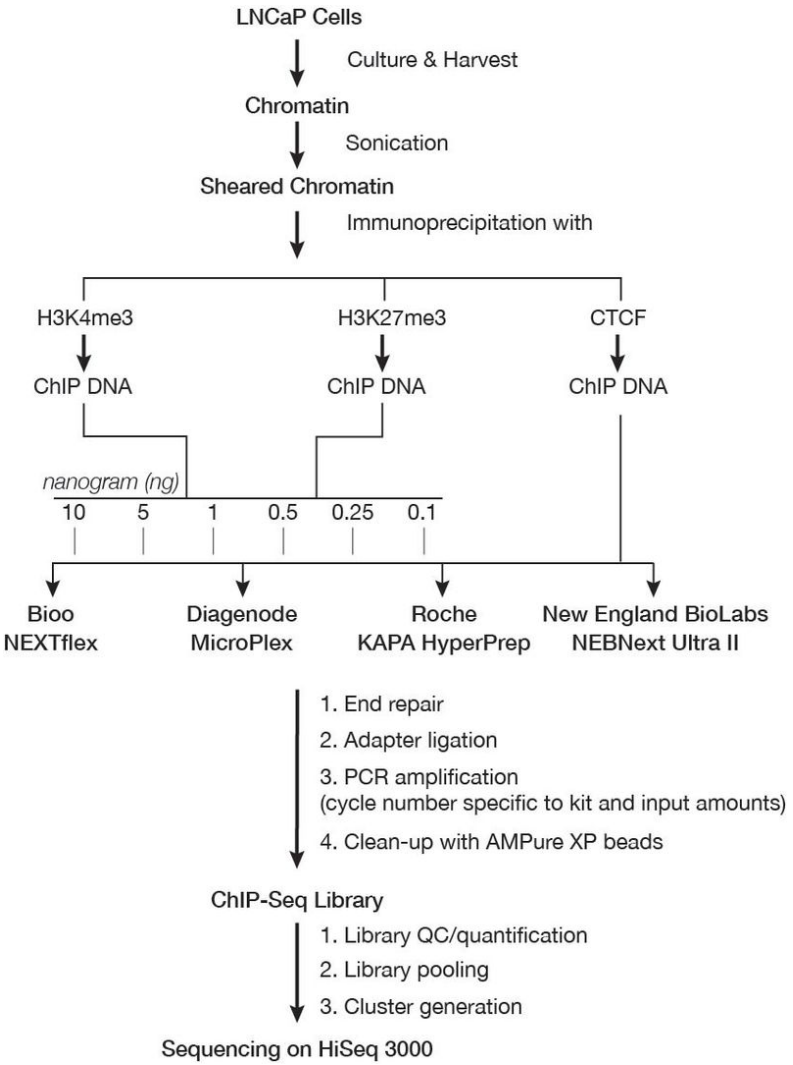

Figure 1

FIGURE 1. Experimental design. Diagram outlining the experimental design for comparing the Bioo NEXTflex, Diagenode MicroPlex, Roche KAPA HyperPrep, and NEB NEBNext Ultra II ChIP-Seq library preparation protocols. Two independent ChIP-Seq experiments (sets 1 and 2) were carried out to assess H3K4me3 and H3K27me3 libraries, and 3 independent ChIP-Seq experiments (sets 1-3) were performed to evaluate CTCF libraries. For H3K4me3 and H3K27me3, 6 different amounts of input DNA (10, 5, and 1 ng and 500, 250, and 100 pg) were used for library construction; however, due to low recovery of CTCF ChIP DNA, only a single DNA input was used for library construction. The number of amplification cycles for each protocol (Table 1) followed the manufacturers' recommendations based on the amount of input DNA.

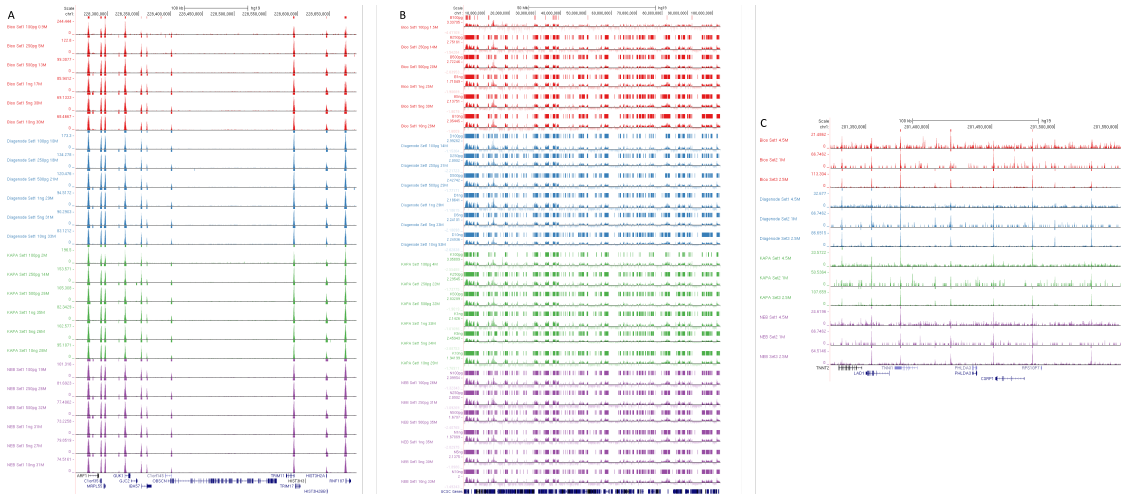

Figure 2

FIGURE 2. Representative screenshots showing the overall signal landscape for each ChIP-Seq library. (*A*) H3K4me3 set 1 libraries. (*B*) H3K27me3 set1 libraries. (*C*) CTCF libraries. The small bars on top of each track indicate the peaks/broad domains called in each library. The number of reads used to generate the signal landscape of each library is indicated at the end of track name.

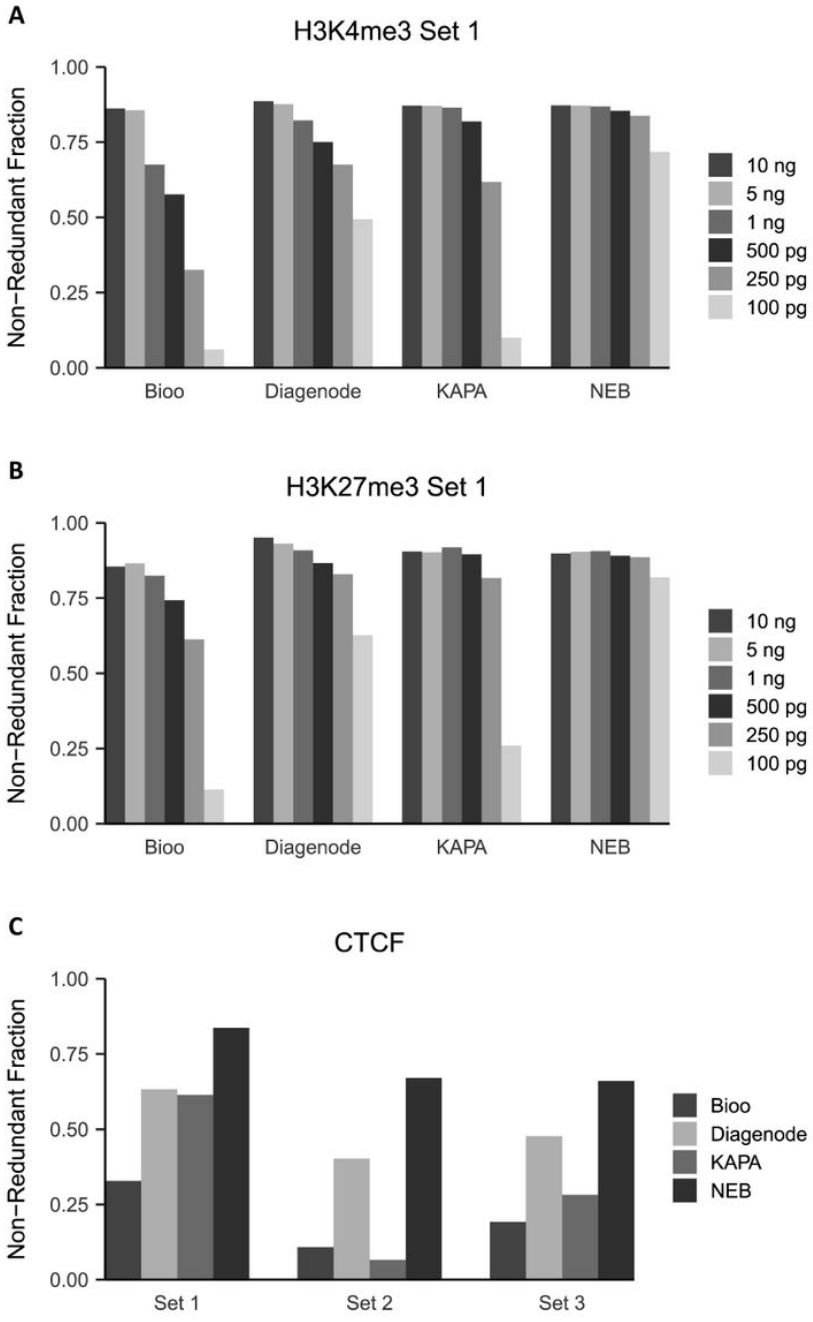

Figure 3

FIGURE 3. Library complexity measured as NRFs. (*A*) NRF for H3K4me3 set 1 libraries. (*B*) NRF for H3K27me3 set 1 libraries. (*C*) NRF for libraries from CTCF sets 1 to 3.

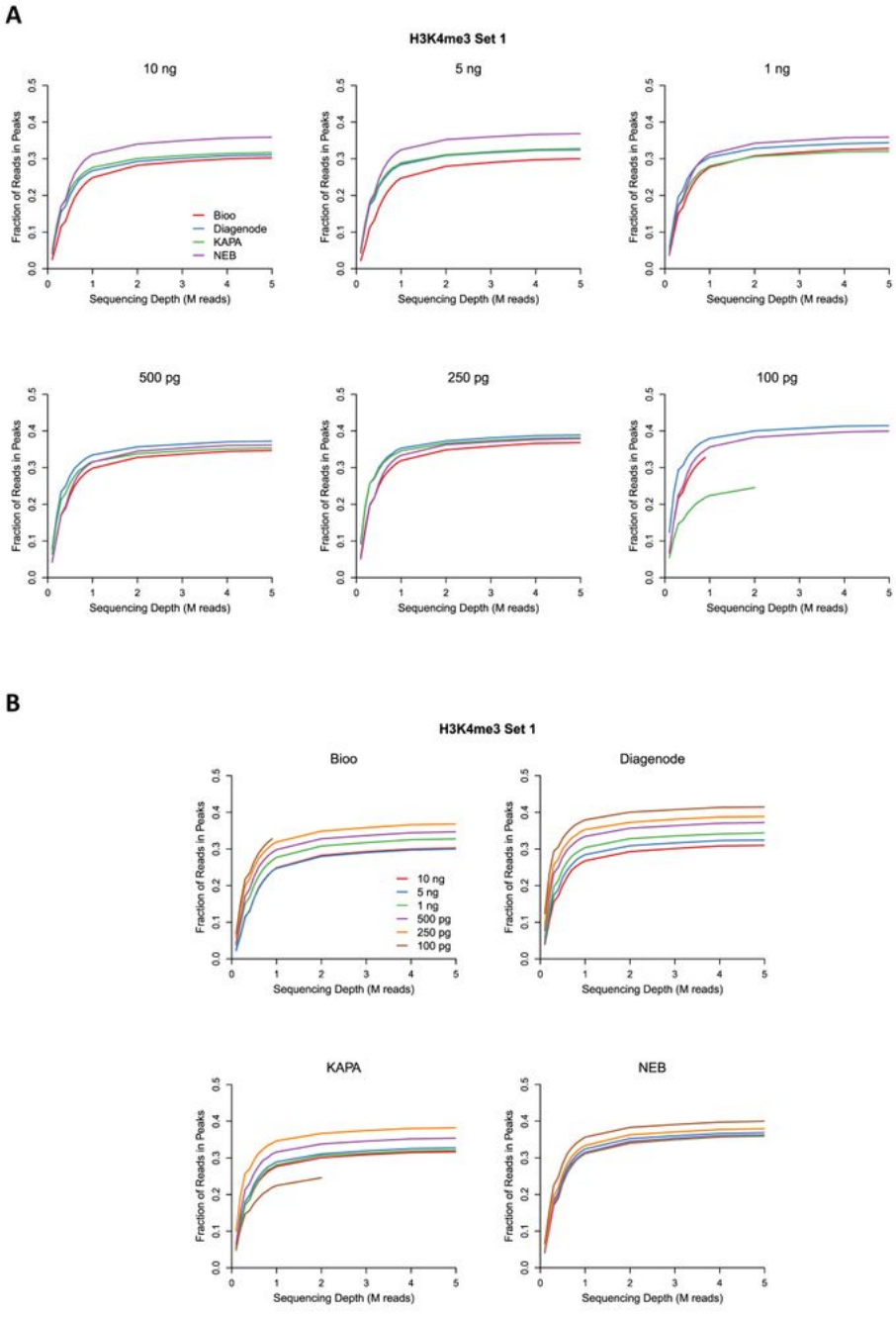

Figure 4

FIGURE 4. FRiP for H3K4me3 plotted against sequencing depth for H3K4me3 set 1 libraries. (*A*) Graphs displayed to highlight differences in protocol performance at specific DNA inputs. (*B*) Graphs displayed to facilitate comparisons of a single protocol at different DNA inputs.

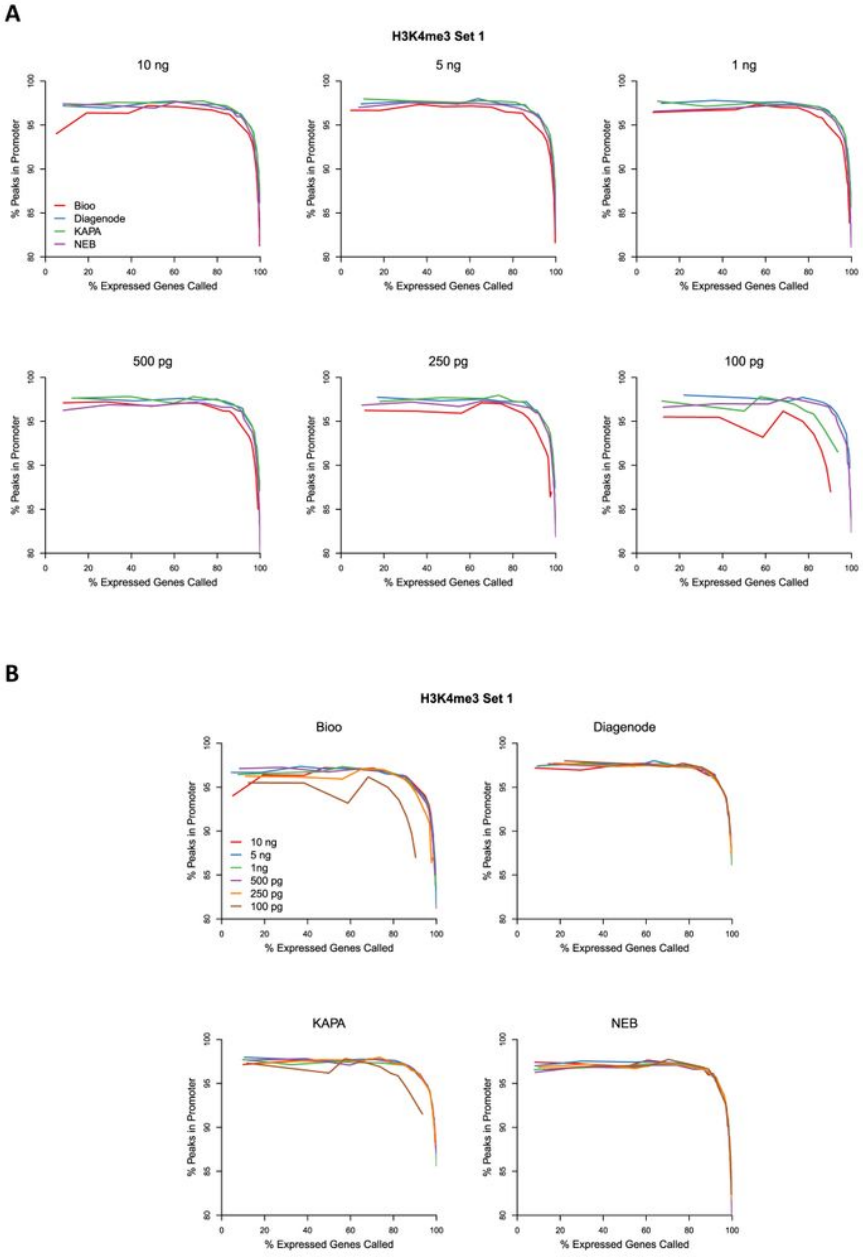

Figure 5

FIGURE 5. The peak calling quality for H3K4me3 measured as the percentage of peaks in promoter regions versus the percentage of expressed genes marked by peaks at varying sequencing depth for H3K4me3 set 1 libraries. (*A*) Graphs displayed to highlight diffences in the performance of each protocol at specific DNA inputs. (*B*) Graphs displayed to facilitate comparisons of a single protocol at different DNA inputs.

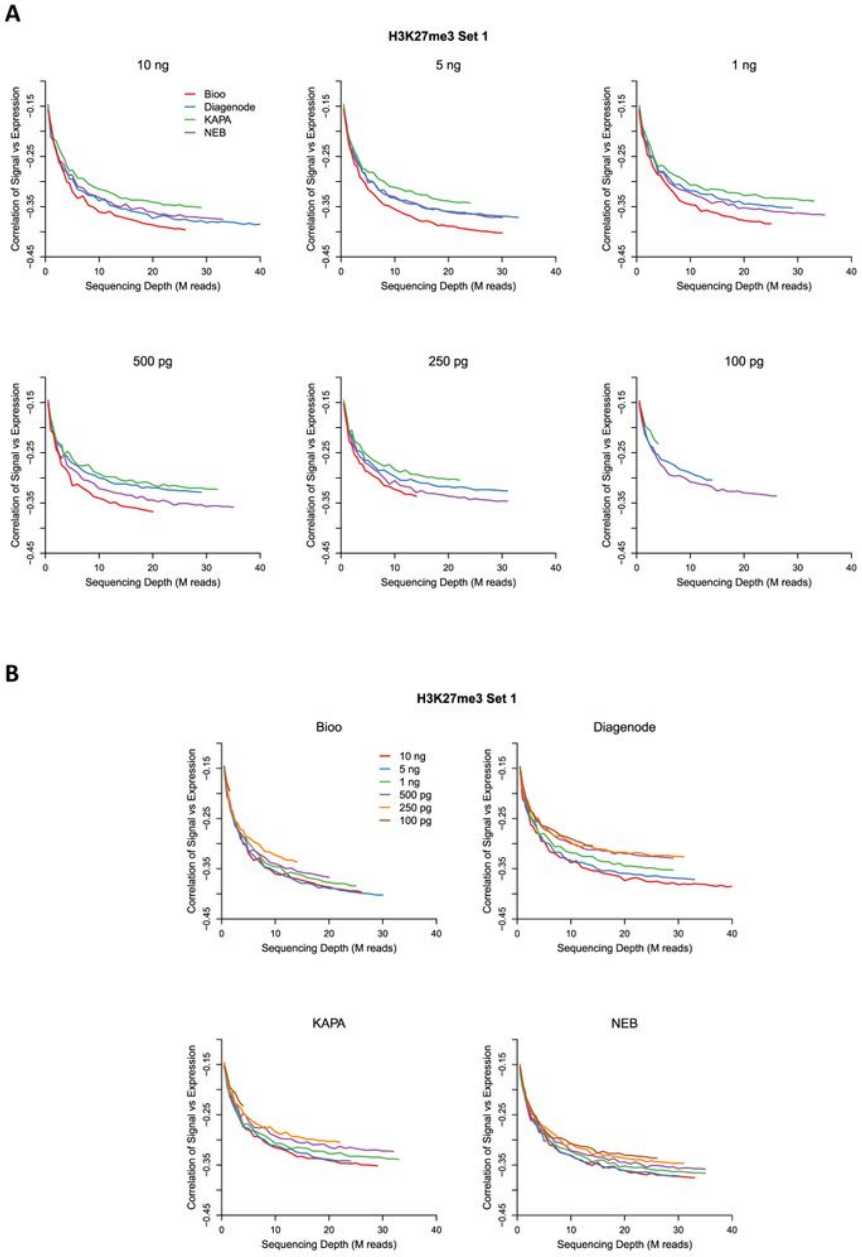

Figure 6

FIGURE 6. Spearman's correlation of H3K27me3 signal intensity versus gene expression plotted against increasing sequencing depth for H3K27me3 set 1 libraries. The lower the curve, the better the negative correlation with gene expression. (*A*) Graphs displayed to highlight diffences in the performance of each protocol at specific DNA inputs. (*B*) Graphs displayed to facilitate comparisons of a single protocol at different DNA inputs.

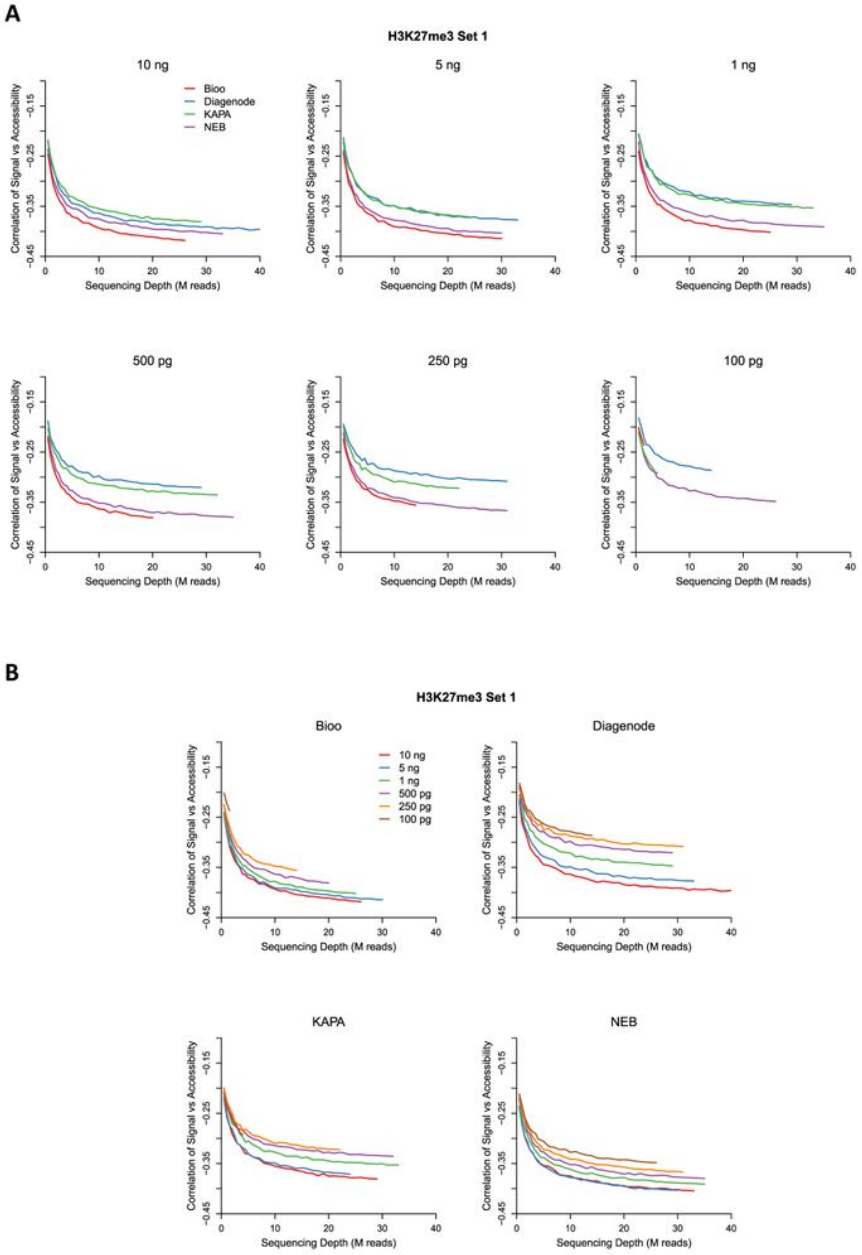

Figure 7

FIGURE 7. Spearman's correlation of H3K27me3 signal intensity versus chromatin accessibility plotted against increasing sequencing depth for H3K27me3 set 1 libraries. The lower the curve, the better the negative correlation with gene expression. (*A*) Graphs displayed to highlight diffences in the performance of each protocol at specific DNA inputs. (*B*) Graphs displayed to facilitate comparisons of a single protocol at different DNA inputs.

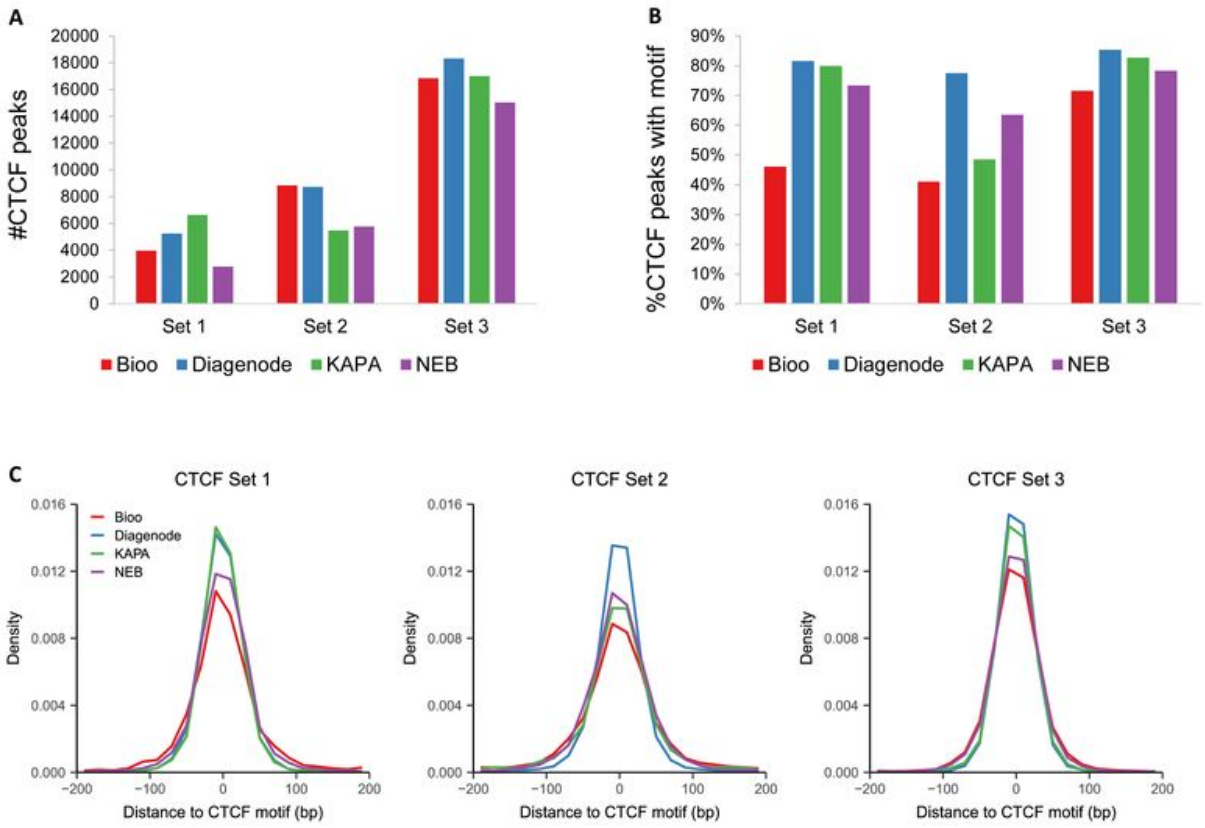

Figure 8

FIGURE 8. Performance of protocols for CTCF ChIP-Seq. Libraries for CTCF sets 1 to 3 are presented. (*A*) Number of peaks called in each library. (*B*) Percentage of peaks with CTCF motifs within 100 bp of peak summits. (*C*) Histogram of distance between peak summit and CTCF motif.

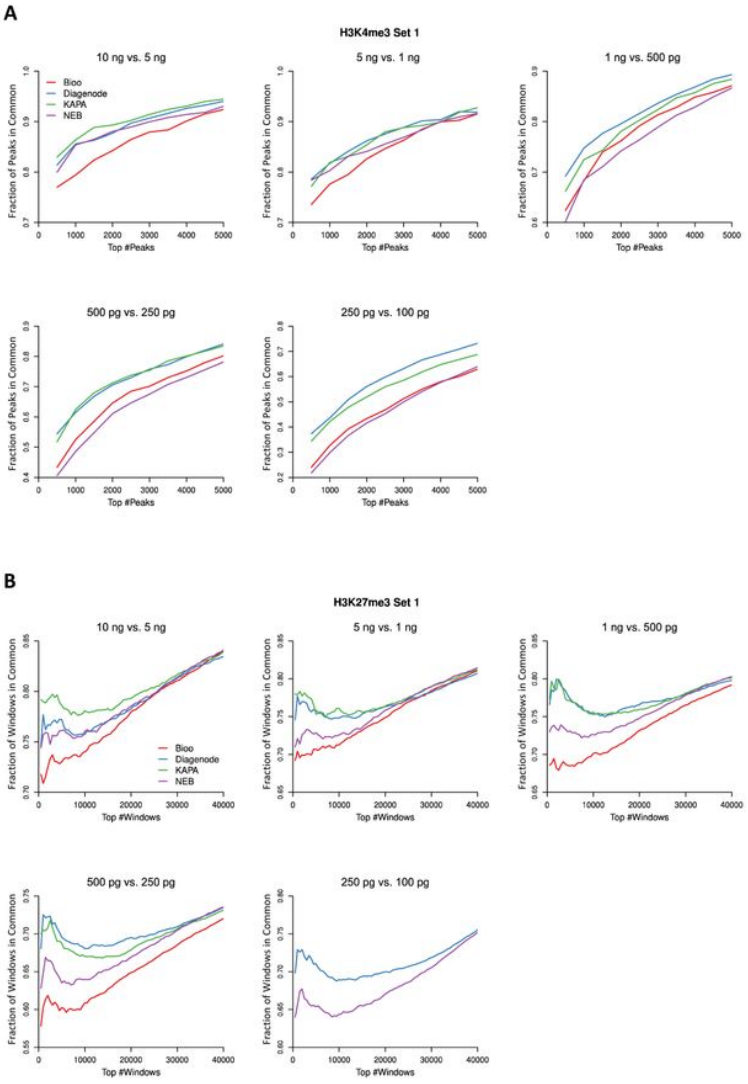

Figure 9

FIGURE 9. Reproducibility curves displayed to highlight differences in the performance of each protocol. The *x*-axis represents the number of top peaks (H3K4me3) or enriched 20-kb windows (H3K27me3) in pairwise comparisions of libraries produced from incremental changes in DNA input (ie, 10 ng versus 5 ng, 5 ng versus 1 ng, 1 ng versus 500 pg, 500 pg versus 250 pg, and 250 pg versus 100 pg). The *y*-axis represents the fraction of top peaks or enriched windows that are in common between the 2 libraries that were compared. H3K4me3 (*A*) and H3K27me3 (*B*) set 1 data are plotted.

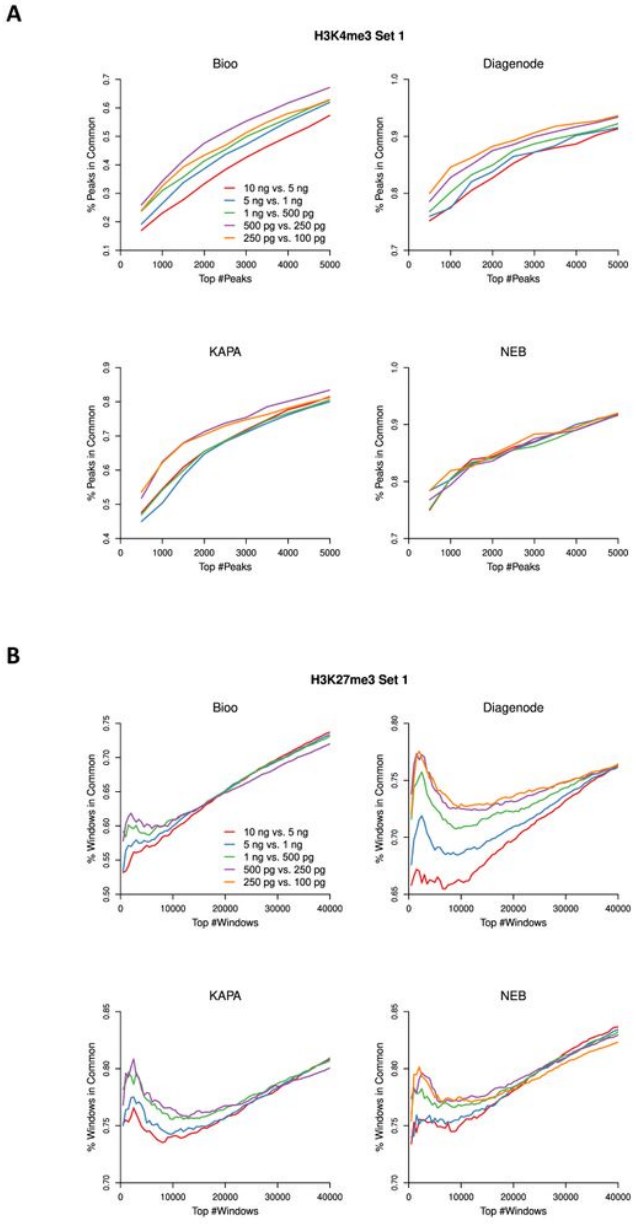

Figure 10

FIGURE 10. Reproducibility curves displayed to facilitate single protocol comparisons at incremental changes in DNA input. The *x*-axis represents the number of top peaks (H3K4me3) or enriched 20-kb windows (H3K27me3) in incremental, pairwise comparisons of DNA input levels. The *y*-axis represents the fraction of top peaks or enriched windows that are in common between the 2 libraries that were compared. H3K4me3 (*A*) and H3K27me3 (*B*) set 1 data are plotted.

Supplemental Material

FIGURE S1. Fraction of H3K4me3 reads in promoter regions for H3K4me3 set 1 libraries. (*A*) Bar graphs displayed to highlight differences in protocol performance at specific DNA inputs. (*B*) Graphs displayed to facilitate comparisons of a single protocol at different DNA inputs.

FigureS1.pdf 64 KB

FIGURE S2. Aggregated signal curves around TSSs for H3K4me3 set 1 libraries organized by protocol.

FigureS2.pdf 58 KB

FIGURE S3. Scatter plots showing H3K4me3 signal intensity in peaks for each amount of input DNA (*y*-axis) compared to 10 ng of input DNA (*x*-axis) for each protocol. H3K4me3 set 1 library data are presented as scatter plots with smoothed color density and with red lines representing locally weighted smoothing (LOESS) fitted curves.

FigureS3.pdf 2 MB

FIGURE S4. Scatter plots showing H3K4me3 signal intensity in peaks for each protocol (*y*-axis) compared to the NEB protocol (*x*-axis) at DNA inputs of 10 and 5 ng. H3K4me3 set 1 library data are presented as scatter plots with smoothed color density and with red lines representing LOESS-fitted curves.

FigureS4.pdf 1 MB

FIGURE S5. Aggregated signal curves around enhancers for H3K4me3 set 1 libraries organized by protocol.

FigureS5.pdf 70 KB

FIGURE S6. Bar graphs indicating the average GC content of the DNA sequences under peaks uniquely called to each protocol at specific DNA inputs. H3K4me3 set 1 libraries are presented in *A*. H3K27me3 set 1 libraries are presented in *B*. H3K27me3 Bioo 100 ng and KAPA 100 ng libraries did not have enough reads for a meaningful peak calling and thus were skipped.

FigureS6.pdf	46 KB

Table S1.

TableS1.xlsx	25 KB

Table S2.

TableS2.xlsx	11 KB

References

1. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell.* 2007;129(4):823-837. ↩
2. Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 2012;22(9):1813-1831. ↩
3. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet.* 2014;15(11):709-721. ↩
4. Gilfillan GD, Hughes T, Sheng Y, et al. Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics.* 2012;13:645. ↩
5. Sundaram AY, Hughes T, Biondi S, et al. A comparative study of ChIP-seq sequencing library preparation methods. *BMC Genomics.* 2016;17(1):816. ↩
6. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57-74. ↩
7. Chao HP, Chen Y, Takata Y, et al. Systematic evaluation of RNA-Seq preparation protocol performance. *BMC Genomics.* 2019;20(1):571. ↩
8. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25. ↩
9. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996-1006. ↩

10. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;47(D1):D76-D773. ↩
11. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14(4):R36. ↩
12. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015;31(2):166-169. ↩
13. Segal T, Salmon-Divon M, Gerlitz G. The heterochromatin landscape in migrating cells and the importance of H3K27me3 for associated transcriptome alterations. *Cells.* 2018;7(11):205. ↩
14. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics.* 2011;27(7):1017-1018. ↩
15. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009;37(suppl_2):W202-W208. ↩
16. Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2020;48(D1):D87-D92. ↩
17. Wan CK, Li P, Spolski R, et al. IL-21-mediated non-canonical pathway for IL-1beta production in conventional dendritic cells. *Nat Commun.* 2015;6:7988. ↩
18. Looney TJ, Zhang L, Chen CH, et al. Systematic mapping of occluded genes by cell fusion reveals prevalence and stability of cis-mediated silencing in somatic cells. *Genome Res.* 2014;24(2):267-280. ↩
19. Manna S, Kim JK, Bauge C, et al. Histone H3 Lysine 27 demethylases Jmjd3 and Utx are required for T-cell differentiation. *Nat Commun.* 2015;6:8152. ↩
20. Hahn MA, Li AX, Wu X, et al. Loss of the polycomb mark from bivalent promoters leads to activation of cancer-promoting genes in colorectal tumors. *Cancer Res.* 2014;74(13):3617-3629. ↩
21. Plasschaert RN, Bartolomei MS. Tissue-specific regulation and function of Grb10 during growth and neuronal commitment. *Proc Natl Acad Sci U S A.* 2015;112(22):6841-6847. ↩
22. Cuadrado A, Remeseiro S, Grana O, Pisano DG, Losada A. The contribution of cohesin-SA1 to gene expression and chromatin architecture in two murine tissues. *Nucleic Acids Res.* 2015;43(6):3056-3067. ↩
23. Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform.* 2017;18(2):279-290. ↩

24. Becker JS, McCarthy RL, Sidoli S, et al. Genomic and proteomic resolution of heterochromatin and its restriction of alternate fate genes. *Mol Cell.* 2017;68(6):1023-1037.e15 ↩

25. Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? *Mol Cell.* 2013;49(5):825-837. ↩

26. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 2012;40(10):e72. ↩

27. Dabney J, Meyer M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques.* 2012;52(2):87-94. ↩

28. Peric-Hupkes D, Meuleman W, Pagie L, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell.* 2010;38(4):603-613. ↩